What Is Claimed:

1. A method for crawling documents comprising:

receiving a uniform resource locator (URL); and

determining whether the URL is associated with a web site that uses session identifiers based, at least in part, on a comparison of a portion of URLs that change between different copies of a document downloaded from the web site.

2. The method of claim 1, wherein the document is a home page of the web site.

3. The method of claim 1, further comprising:

extracting, when the URL is associated with a web site that uses session identifiers, a session identifier from the URL to obtain a clean URL; and

determining whether the URL has already been crawled based, at least in part, on a comparison of the clean URL to a set of clean URLs that represent previously crawled URLs.

4. The method of claim 3, wherein the portion of the URLs that change are identified using URLs that are local to the web site.

5. The method of claim 3, wherein the session identifiers from the URLs are extracted using rules for the web site.

6. The method of claim 5, wherein the rules are determined automatically.

7. The method of claim 3, further comprising:

receiving the URL as a URL from a previously crawled web document.

8. The method of claim 3, further comprising:

crawling the URL when the URL is determined to not already have been crawled.

9. The method of claim 1, wherein the comparison determines that the web site uses session identifiers when the portion of the URLs that change is greater than a predetermined value.

10. A method for identifying web sites that use session identifiers comprising:

downloading at least two different copies of at least one document from a web site;

extracting uniform resource locators (URLs) from the two different copies of the web document;

comparing the extracted URLs of the two different copies of the document; and

determining whether the web site uses session identifiers based on the comparison.

11. The method of claim 10, wherein the determining whether the web site uses session identifiers further includes:

determining that the web site uses session identifiers when the comparison indicates that at least a predetermined portion of the URLs change between the two different copies.

12. The method of claim 10, wherein extracting URLs from the two different copies of the document includes extracting only URLs that are local to the web site.

13. The method of claim 10, wherein the document is a home page of the web site.

14. The method of claim 10, further comprising:

analyzing the extracted URLs, when the web site is determined to use session identifiers, to generate at least one rule identifying how the session identifiers are embedded in the URLs.

15. A device comprising:

a spider component configured to crawl web documents associated with at least one web site; and

a session identifier component configured to determine whether the web site uses session identifiers based on a comparison of a portion of uniform resource locators (URLs) that change between different copies of at least one web document downloaded from the web site.

16. The device of claim 15, wherein the spider component further comprises:

at least one fetch component configured to download content from a network; and

a content manager configured to extract URLs from the downloaded content.

17. The device of claim 16, wherein the spider component further comprises:

a URL manager configured to store the extracted URLs.

18. The device of claim 15, wherein the at least one web document is a home page of the web site.

19. The device of claim 15, wherein the portion of the URLs that change are identified from URLs that are local to the web site.

20. The device of claim 15, further comprising:

a session rule generator configured to generate rules describing how the web site embeds session identifiers in the at least one web document.

21. A device comprising:

means for downloading at least two different copies of at least one web document from a web site;

means for extracting uniform resource locators (URLs) from the two different copies of the web document;

means for comparing the extracted URLs of the two different copies of the web document; and

means for determining whether the web site uses session identifiers based on the comparison.

22. The device of claim 21, wherein the means for determining determines that the web site uses session identifiers when the comparison indicates that at least a predetermined portion of the URLs change between the two different copies.

23. The device of claim 21, wherein the means for extracting URLs from the two different copies of the web document includes means for extracting only URLs that are local to the web site.

24. The device of claim 21, wherein the web document is a home page of the web site.

25. The device of claim 21, further comprising:

means for analyzing the extracted URLs, when the web site is determined to use session identifiers, to generate rules describing how the session identifiers are embedded in the URLs.

26. A computer-readable medium containing programming instructions that when executed by at least one processor cause the processor to perform a method for identifying web sites that use session identifiers including:

downloading at least two different copies of at least one document from a web site;

extracting uniform resource locators (URLs) from the two different copies of the document;

comparing the extracted URLs of the two different copies of the web document; and

determining whether the web site uses session identifiers based on the comparison.

27. The computer-readable medium of claim 26, wherein the determining when the web site uses session identifiers further includes:

determining that the web site uses session identifiers when the comparison indicates that at least a predetermined portion of the URLs change between the two different copies.

28. The computer-readable medium of claim 26, wherein extracting URLs from the two different copies of the web document includes extracting only URLs that are local to the web site.

29. The computer-readable medium of claim 26, wherein the web document is a home page of the web site.

30. The computer-readable medium of claim 26, further comprising instructions that cause the at least one processor to:

analyze the extracted URLs, when the web site is determined to use session identifiers, to generate at least one rule describing how the session identifiers are embedded in the URLs.

31. A method for crawling documents comprising:

receiving a uniform resource locator (URL); and

determining whether the URL is associated with a web site that uses session identifiers based, at least in part, on a comparison of content between different duplicate or near-duplicate copies of a document downloaded from the web site for two different URLs.